Measuring the Power of Learning.®

Research Report
ETS RR–16-01

An Evaluation of the Single-Group Growth Model as an Alternative to Common-Item Equating

Youhua Wei

Rick Morgan

June 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

An Evaluation of the Single-Group Growth Model as an Alternative to Common-Item Equating

Youhua Wei & Rick Morgan

Educational Testing Service, Princeton, NJ

As an alternative to common-item equating when common items do not function as expected, the single-group growth model (SGGM) scaling uses common examinees or repeaters to link test scores on different forms. The SGGM scaling assumes that, for repeaters taking adjacent administrations, the conditional distribution of scale scores in later administration, given the scale score in earlier administration, should generalize from previous repeaters to the repeaters taking the current administration. The current repeaters' scale score distribution is estimated from their earlier scale score distribution. The SGGM scaling first uses previous repeaters' data to estimate the conditional distribution of their later scale scores, given their earlier scale scores. Then the repeaters taking both the current and previous administrations are identified, and their scale score distribution on the current form is estimated based on their previous scale score distribution and the estimated conditional distribution. Finally, a single-group equipercentile equating is performed between the current-form repeaters' observed raw score distribution and their estimated scale score distribution to obtain the raw-to-scale score conversion. This study evaluated the SGGM scaling performance using the common-item equating results for a language test as the criterion. The study found that the raw-to-scale conversions based on SGGM scaling differed from those based on common-item equating. However, the SGGM scaling results did not show a systematic bias in either the average or the variability of examinees' scale scores.

Keywords Repeater; single-group growth model; linking

doi:10.1002/ets2.12087

The common-item equating design, specifically the nonequivalent groups with anchor test (NEAT) design, is a widely used data collection design in test score equating practice because of its flexibility in operational work. However, the administrative flexibility of this design is obtained at the cost of a lack of empirical data to examine the assumptions of NEAT equating methods (Kolen & Brennan, 2004) and the potential risk of common-item exposure. The common-item exposure risk becomes more serious for a high-stakes test in a population where test takers are able to share the test items on the Internet after they take the test. Exposure to common items positively affects examinee performance, and as a result, the NEAT design results in overestimating the achievement levels of examinees, leading to less than optimal score equating. New alternative designs without using common items need to be explored and investigated for test equating in security-risk populations. This study evaluated one of the proposed designs and methods—single-group growth model (SGGM) scaling. The data used for the evaluation came from a number of administrations of a secured English-language test, and NEAT design equating results were used as the criterion.

To address the common-item security issue, six alternative methods have been explored as substitutes for NEAT design equating. Liao and Livingston (2012) explored three approaches. The first approach eliminated equating by trying to create randomly equivalent forms. The results showed that the form-to-form differences were too large for the technique to be an adequate substitute for score equating. The second approach created equivalent groups using demographic information so that the new and reference forms could be directly equated as in equivalent-groups equating design. The approach did not fully adjust the ability of the new-form group to be similar enough to that of the reference form group and therefore was inadequate as a substitute for common-item equating. The third approach was a single-group equating design that treated the group of examinees who took two forms on adjacent administrations 1 month apart as being the same group. The operational data from an English-language test showed score gains for the 1-month repeaters, thus not confirming the assumption of equal ability of the adjacent-month repeaters. Therefore, the single-group scaling approach could not be used as a substitute for common-item equating. However, Livingston (2005) proposed an adjusted single-group scaling

Corresponding author: Y. Wei, E-mail: YWei@ets.org

method for a language test, using the scores of people repeating the test with only 1 month between the two administrations. People who were taking the test for the first time at the earlier administration were to be excluded from the analysis.

Wei, Liu, and Dorans (2013) explored and evaluated the fourth and fifth alternative NEAT linking designs. One was to conduct equating in a population where common-item exposure was not a concern and then apply the same conversion to the population where the security of common items was a big concern. The other alternative was to conduct linking across two different populations, where a new form administered to the high security-risk population was linked to an old form administered to the low security-risk population through common items. They compared results from these two alternative linking designs with those from the linking conducted within the high security-risk population, which was considered as the criterion linking function. The findings suggested that the linking functions from both alternative linking designs were different from the criterion linking function, although the cross-population linking results were closer to the criterion function.

Morgan (2011) proposed the sixth alternative method: the SGGM scaling method. With the consideration of the adjacent-month repeaters' score gains, this method is a further development of the single-group scaling approach explored by Liao and Livingston (2012) and the adjusted single-group scaling proposed by Livingston (2005). The SGGM scaling borrows the statistical procedure used in frequency estimation equating, where the conditional distribution of total score given each anchor score in one population is assumed the same as in another population. In SGGM scaling, the conditional distribution of adjacent-administration repeaters' later scale scores given each earlier scale score based on historical data is assumed to be consistent across administrations, so it can be used to estimate the current administration repeaters' scale score distribution on the new test form. Specifically, SGGM scaling comprises three steps:

- *Establish repeaters' scale score conditional distribution based on historical data.* First, compile historical data for examinees taking the two different test forms in adjacent administrations (e.g., repeaters taking the test a month apart). Then use the data to estimate the conditional distribution of those repeaters' scale scores in later administration, given their scale scores in earlier administration. Estimate this conditional distribution for each scale score of the earlier administration. Following the notation used in frequency estimation equating (Kolen & Brennan, 2004), the conditional distribution of repeaters' later scale score X_2, given their earlier scale score X_1, can be defined as $f_{\text{old}}(x_2|x_1)$, which represents the probability that $X_2 = x_2$ given that $X_1 = x_1$ based on old or historical data.
- *Calculate the expected scale score distribution for repeaters in the new administration.* Among the examinees taking the new test form in the current administration, identify those who took the test form in the previous administration. Then use their scale score distribution from the previous administration, along with the conditional distribution estimated in Step 1, to estimate their scale score distribution in the current administration. Assuming the new repeaters' scale score Y_1 in the previous administration has distribution $g_{\text{new}}(y_1)$, the repeaters' expected scale score distribution in the current administration, $g_{\text{new}}(y_2)$, can be obtained by

$$g_{\text{new}}(y_2) = \sum_{y1} f_{\text{old}}(x_2|x_1)\, g_{\text{new}}(y_1).$$

- *Conduct raw-to-scale score scaling for the new test form.* In the group of test takers in Step 2 (i.e., the new adjacent-administration repeaters), perform a single-group equipercentile equating of their new-form raw score distribution $h_{\text{new}}(z)$ to their scale score distribution $g_{\text{new}}(y_2)$ estimated in Step 2. Then use the resulting transformation as the raw-to-scale score conversion for the new test form. For all examinees in the new administration, use the conversion for score reporting.

The critical step in SGGM scaling is estimating the expected scale score distribution for repeaters in the new administration based on historical data. To illustrate the process, Tables 1 – 4 set forth a hypothetical example. Assume we have a test with five possible raw score points 0, 1, 2, 3, and 4 and three possible scale score points 10, 20, and 30. Table 1 shows the bivariate frequency distribution of earlier and later scale scores of repeaters who have taken two different test forms in adjacent administrations in the past. The table shows the frequency of examinees at each earlier scale score level. For example, 200 repeaters received a score of 10 in the earlier administration. Among those repeaters, 80 received a score of 10, 100 received a score of 20, and 20 received a score of 30 in the later administration. From the frequency distribution, we can compute the conditional distribution of repeaters' later scale scores given each of their earlier scale scores. For

Table 1 Historical Repeaters' Scale Score Frequency Distribution

		Later score X_2			
		10	20	30	Sum
Earlier score X_1	10	80	100	20	200
	20	100	200	100	400
	30	20	40	140	200

Table 2 Historical Repeaters' Scale Score Conditional Distribution

		Later score X_2		
		10	20	30
Earlier score X_1	10	0.4 (80/200)	0.5 (100/200)	0.1 (20/200)
	20	0.25 (100/400)	0.50 (200/400)	0.25 (100/400)
	30	0.1 (20/200)	0.2 (40/200)	0.7 (140/200)

Table 3 New Repeaters' Scale Score Frequency Distribution

		Later score Y_2			
		10	20	30	Sum
Earlier score Y_1	10	4 (10×0.4)	5 (10×0.5)	1 (10×0.1)	10
	20	15 (60×0.25)	30 (60×0.5)	15 (60×0.25)	60
	30	2 (20×0.1)	4 (20×0.2)	14 (20×0.7)	20
Sum		21 ($4+15+2$)	39 ($5+30+4$)	30 ($1+15+14$)	

example, for repeaters with score 10 in earlier administrations, the probabilities of receiving scores 10, 20, and 30 in the later administration are $80/200 = 0.4$, $100/200 = 0.5$, and $20/200 = 0.1$. Table 2 shows repeaters' conditional distribution of later scores at each earlier score. For the new administration, we can identify repeaters who took the test in the previous administration and obtain their earlier scale scores. Table 3 shows the frequency of repeaters at each earlier score level. There are 10 scores of 10, 60 scores of 20, and 20 scores of 30. We can estimate the bivariate frequency distribution of new repeaters' earlier and later scale scores by applying the conditional probabilities in Table 2 to the frequencies of Table 3. For example, among the 10 repeaters with earlier score 10, four ($10 \cdot 0.4$), five ($10 \cdot 0.5$), and one ($10 \cdot 0.1$) repeaters will receive scores 10, 20, and 30, respectively, in the later administration. Table 3 shows the estimated bivariate frequency distribution of new repeaters' earlier and later scale scores. Then we can add up the frequencies in each column and obtain the estimated frequency at each later score level. Specifically, there are 21, 39, and 30 repeaters with scores 10, 20, and 30, respectively, in the new administration. Table 4 shows the repeaters' expected scale score frequency distribution. The left two columns show their observed raw score frequency distribution. Then we can do a single-group equipercentile equating of repeaters' raw score distribution to their scale score distribution.

Wei (2013) compared the SGGM scaling and the cross-population NEAT linking with the single-group equating method, which assumes equal ability of repeaters taking the test within 2 weeks. Although the results from both alternative linking methods were different than those from the single-group equating, the SGGM scaling results were more similar to the single-group equating results, and the cross-population NEAT linking method tended to produce higher conversion lines, higher scale score means, and more repeaters' average score changes than the single-group equating did.

Based on previous studies, although SGGM scaling has the assumption of consistency of repeaters' score gains over administrations, it has an attractive potential to link test scores in a population where common items are not secure to appear in any different test form or do not function as expected. The main purpose of this study was to examine the assumption of the SGGM scaling method and compare its linking results with those from the well-designed NEAT equating method using the data collected from a secured language-testing program.

Table 4 New Repeaters' Raw and Scale Score Distributions

Raw score distribution		Scale score distribution	
Raw score	Frequency	Scale score	Frequency
0	1	10	21
1	13	20	39
2	30	30	30
3	39		
4	9		

Methodology

This study used operational data from a testing program to evaluate the performance of the SGGM scaling method. We first used repeaters' scale scores in adjacent administrations collected earlier in the testing program to estimate the conditional distribution of repeaters' later scale scores, given their earlier scores. We then estimated new repeaters' scale score distribution in the current administration based on their scale scores in the earlier administration. We conducted single-group equipercentile equating between the current-form repeaters' observed raw score distribution and their estimated scale score distribution to obtain the raw-to-scale score conversion. We compared the SGGM scaling results with the NEAT equating results to examine the performance of the SGGM scaling method.

Data

This study used data from a language test designed to measure examinees' English proficiency. The test consists of two separately timed and scored sections: listening and reading. Each section comprises 100 multiple-choice items. A well-designed NEAT equating is conducted separately for the two sections. The testing program closely monitors the security of common items. The scaled scores are reported on a scale ranging from 5 to 495, by increments of 5, for each section.

The study utilized data from the listening and reading test scores collected from May 2010 to November 2012 administrations (with each administration having one test form). We used data from all examinees taking adjacent administrations in the selected time period, either 1 month or 2 months apart, in the analyses for this study. Table 5 shows the adjacent administrations and repeaters' sample sizes. It is likely that some examinees appeared in more than one of these repeater groups because they may have taken the test multiple times within the time of period. The scale score data of 14 groups of

Table 5 Summary of Adjacent-Administration Repeaters' Data

	Administrations for conditional distribution		Administrations for scaling	
Time gap	Adjacent administrations	Repeater *N*	Adjacent administrations[a]	Repeater *N*
1 month	May 10 – June 10	9,188	May 12 – June 12 (A)	6,449
	June 10 – July 10	6,251	May 12 – June 12 (B)	5,763
	September 10 – October 10	12,666	June 12 – July 12 (C)	4,533
	October 10 – November 10	14,598	June 12 – July 12 (D)	4,020
	May 11 – June 11	11,972	September 12 – October 12 (G)	4,096
	June 11 – July 11	7,935	September 12 – October 12 (H)	2,728
	September 11 – October 11	15,334	October 12 – November 12 (I)	4,056
	October 11 – November 11	15,920	October 12 – November 12 (J)	5,782
	Cumulative	93,864		
2 months	July 10 – September 10	9,979	July 12 – September 12 (E)	3,696
	November 10 – January 11	14,405	July 12 – September 12 (F)	2,088
	July 11 – September 11	13,033		
	November 11 – January 12	16,971		
	January 12 – March 12	19,735		
	March 12 – May 12	22,491		
	Cumulative	96,614		
Combined	Cumulative	190,478		

[a] A – F indicate the test forms administered in later administrations.

repeaters from May 2010 to May 2012 were used to estimate the conditional distribution of later scale scores, given earlier scale scores. We also examined the consistency of scale score gains across different pairs of adjacent administrations. The SGGM scaling used the data collected from June 2012 to November 2012 administrations. Specifically, for each pair of adjacent administrations, we used repeaters' scale scores in the earlier administration and raw scores in the later administration to perform SGGM scaling. We compared the SGGM scaling results with the NEAT equating results used in the operational work. The administrations used to estimate conditional distribution included eight pairs of adjacent administrations with a time gap of 1 month and six pairs with a time gap of 2 months. The repeater sample sizes ranged from 6,251 to 22,491. A total of 190,478 pairs of scale scores was used to estimate and examine the conditional distribution. The administrations used for scaling included eight pairs of adjacent administrations with a time gap of 1 month and two pairs with a time gap of 2 months. The repeater sample sizes ranged from 2,088 to 6,449 for the score scaling.

Procedure

Computing Conditional Distribution

To evaluate the assumption of SGGM scaling (i.e., the consistency of repeaters' score gains across different pairs of adjacent administrations), we computed the conditional distribution of repeaters' later scale scores, given each of the earlier scale scores, four different ways. The conditional distributions were based on (a) the data of each pair of adjacent administrations from May 2010 to May 2012, (b) the accumulated data of the eight pairs of adjacent administrations with the time gap of 1 month, (c) the accumulated data of the six pairs of adjacent administrations with the time gap of 2 months, and (d) the accumulated data of the 14 pairs of adjacent administrations with the time gap of 1 or 2 months.

Examining the Consistency of Score Gains

For each of the adjacent-administration repeater groups from May 2010 to May 2012, we first computed the mean and standard deviation of their score gains then computed the mean and standard deviation of their score gains at each scale score point on early administration. We then compared these means and standard deviations of score gains across different repeater groups in (a) the eight pairs of adjacent administrations with the time gap of 1 month, (b) the six pairs of adjacent administrations with the time gap of 2 months, and (c) the accumulated pairs with 1-month and 2-month gaps. The purpose was to evaluate the consistency of repeaters' score gains across administrations and then select the conditional score distribution for SGGM scaling.

Selecting the Conditional Distribution

Based on the comparison of repeaters' score gains with a 1-month gap versus the repeaters' score gains with a 2-month gap, we made a decision about which conditional distribution would be used for scaling in the next step. Before creating the final conditional distribution for scaling, we smoothed the repeaters' scale score distributions from earlier and later administrations using a bivariate log linear model that preserved 16 moments of the observed bivariate distribution: the first six univariate moments of each variable and the first four bivariate moments (i.e., X_1X_2, $X_1^2X_2$, $X_1X_2^2$, and $X_1^2X_2^2$; Holland & Thayer, 2000).

Conducting Scaling

Based on NEAT equating results used in the operational work, for each of the adjacent-administration repeater groups from June 2012 to November 2012, we used the scale score distribution in the earlier administration to estimate the scale score distribution in the later administration using the procedure described earlier in this report. We then performed a single-group equipercentile equating between these repeaters' raw score distribution on the new form and their estimated scale score distribution on the new administration. The resulting raw-to-scale conversion was used to produce scale scores for all examinees who had taken the new test form. Before conducting the scaling, we smoothed the scale score distribution from the earlier administration and the observed raw score distribution from the new administration using a univariate log linear model that preserved the first six moments of the observed univariate distribution.

Evaluating Scaling Results

We compared the SGGM scaling results with the NEAT equating results. NEAT equating worked well in the testing program, as there were typically small ability differences between the new and reference groups, and the anchor items performed consistently in the new and reference test forms. We examined the raw-to-scale score conversions of the test forms based on the SGGM method. Specifically, for each form used in the later administrations from June 2012 to November 2012, we first computed the raw-to-scale conversion differences by subtracting the scale score based on NEAT equating from the scale score based on SGGM scaling at each raw score point then plotted the scale score differences against the new-form raw scores. We used the raw-to-scale conversion difference plots to evaluate the accuracy of the SGGM scaling method for each of the 10 test forms. To summarize the raw-to-scale conversion differences and examine the SGGM scaling function across the 10 test forms, we computed the root-mean-square of differences (RMSD) between SGGM and NEAT conversions at each raw score point x by

$$\mathrm{RMSD} = \sqrt{\frac{1}{10}\sum_{1}^{10}\left[\mathrm{SS}_{\mathrm{SGGM}}(x) - \mathrm{SS}_{\mathrm{NEAT}}(x)\right]^2},$$

where $\mathrm{SS}_{\mathrm{SGGM}}(x)$ is the scale score from SGGM scaling conversion and $\mathrm{SS}_{\mathrm{NEAT}}(x)$ is the scale score from NEAT equating conversion for raw score x.

We also evaluated the scoring results of the SGGM method for test takers in each of the 10 administrations. Specifically, we first applied the raw-to-scale conversions from the SGGM scaling to all test takers in each administration and obtained their scale scores then computed the means and standard deviations of their scale scores. We compared these means and standard deviations with those from NEAT equating.

To examine the impact of the conversion differences on examinees' scale scores across the scale score range, we computed the weighted RMSD (WRMSD) for each test form by

$$\mathrm{WRMSD} = \sqrt{\frac{\sum_{x} w_x\left[\mathrm{SS}_{\mathrm{SGGM}}(x) - \mathrm{SS}_{\mathrm{NEAT}}(x)\right]^2}{\sum_{x} w_x}},$$

where w_x is the frequency at the raw score point x.

Results

We first compared different repeater groups' score gains and examined the consistency of their score gains across different pairs of adjacent administrations. Then we compared the raw-to-scale conversions and examinees' scale score means and standard deviations based on the SGGM scaling method with those based on the NEAT equating method. We also used WRMSD for the comparison.

Consistency of Score Gains

We first computed scale score gains for each repeater group across all scale score levels and then computed score gains for subgroups at each scale score level for adjacent administrations with a 1-month gap, a 2-month gap, and the combined time gaps. We present the results separately for the listening and reading sections.

Listening

Table 6 shows the adjacent-administration repeaters' scale score means in earlier and later administrations and the means and standard deviations of their score gains. For the eight pairs of adjacent administrations with a 1-month gap, the average score gains ranged from 2.73 to 11.58, with standard deviations from 39.12 to 42.16. Combining the eight pairs of adjacent administrations, the average score gain was 6.63, with a standard deviation of 41.21. For the six pairs of adjacent administrations with a 2-month gap, the average score gains ranged from 4.17 to 12.40, with standard deviations from

Table 6 Summary Statistics of Repeaters' Listening Scale Score Gains

			Earlier admin.		Later admin.		Score gain		
Time gap	Adjacent administration	Sample size	*M*	*SD*	*M*	*SD*	*M*	*SD*	Standardized difference
1 month	May 10 – June 10	9,188	317.75	78.26	329.33	79.48	11.58	42.16	0.15
	June 10 – July 10	6,251	332.69	81.08	335.43	80.54	2.73	40.30	0.03
	September 10 – October 10	12,666	329.30	80.70	337.01	78.37	7.71	40.76	0.10
	October 10 – November 10	14,598	327.01	77.55	331.72	75.77	4.71	39.12	0.06
	May 11 – June 11	11,972	321.64	76.21	330.64	78.88	9.00	41.81	0.12
	June 11 – July 11	7,935	332.96	80.74	337.88	78.65	4.92	41.07	0.06
	September 11 – October 11	15,334	335.55	79.80	341.76	79.40	6.21	42.13	0.08
	October 11 – November 11	15,920	332.69	78.95	338.36	76.26	5.67	41.60	0.07
	Accumulative	93,864	328.97	79.19	335.60	78.24	6.63	41.21	0.08
2 months	July 10 – September 10	9,979	331.07	79.92	335.24	81.75	4.17	41.34	0.05
	November 10 – January 11	14,405	321.87	76.13	334.27	76.97	12.40	41.09	0.16
	July 11 – September 11	13,033	333.49	79.15	339.76	81.46	6.27	42.32	0.08
	November 11 – January 12	16,971	329.26	76.64	338.21	77.91	8.95	40.51	0.12
	January 12 – March 12	19,735	329.40	79.73	336.42	77.91	7.03	40.79	0.09
	March 12 – May 12	22,491	325.53	76.97	331.65	76.55	6.12	41.95	0.08
	Accumulative	96,614	328.08	78.04	335.63	78.40	7.56	41.39	0.10
1 and 2 months	Accumulative	190,478	328.51	78.61	335.61	78.32	7.10	41.31	0.09

40.51 to 42.32. Combining the six pairs of administrations, the average score gain was 7.56, with a standard deviation of 41.39. For the 14 pairs of adjacent administrations, the standardized mean score changes varied from 0.03 to 0.16 for different repeater groups. Therefore, some variability was observed in the average score gains across different pairs of adjacent administrations. Although we found considerable variability in examinees' score gains, the dispersion of score gains was very consistent across different pairs of adjacent administrations. Both the average and the variability of score gains of the accumulated 1-month repeaters and 2-month repeaters were very similar, and their combined average score gain was 7.10, with a standard deviation of 41.31.

To explore in more detail the variability of score gains across score ranges, we examined the conditional means and standard deviations of score gains at each scale score points for both separate pairs and accumulated pairs of adjacent-administration repeaters. Figures 1a and b show the conditional means of score gains for different pairs of administrations with a 1-month gap and a 2-month gap, Figure 1c shows the conditional means of score gains based separately on accumulated 1-month and 2-month data sets, and Figure 1d shows the conditional means of score gains based on combined and accumulated data sets. Figure 2 shows the conditional standard deviations of score gains based on different data sets. In these figures, the horizontal axis indicates the earlier administration's scale score and the vertical axis indicates the conditional mean or standard deviation of scale score gains for each early scale score. The conditional statistics for scale scores below 115, which approximately corresponds to the chance raw score level of 28, were not presented in the figures because very few examinees had scores in that range. Figure 2a shows the conditional standard deviations of score gains for Listening with a 1-month gap; Figure 2b, with 2-month gap; Figure 2c, with 1- and 2-month gaps; and Figure 2d, with a 1- or 2-month gap.

From Figures 1 and 2, one can see that the score gains were different across different scale score points. For examinees with earlier scale scores higher than 400, scale scores tended to drop, and the amount of score drop was related to the examinees' earlier score levels: The higher their earlier scores were, the more scores dropped later. For examinees with earlier scores lower than 400, scale scores tended to increase, and the amount of score increase was also related to the examinees' earlier score levels: The lower their earlier scores were, the more their scores increased later. The variability of score gains was also related to repeaters' earlier score levels: The dispersion of score gains tended to be relatively smaller at the higher score level (higher than 400) and bigger at the lower score level (lower than 200), although there was variability across pairs of administrations. At each scale score point, there was some variability in both conditional means and standard deviations of score gains across different pairs of adjacent administrations with either 1- or 2-month gaps. However, the conditional statistics of score gains over a 1-month interval between adjacent administrations were similar to those

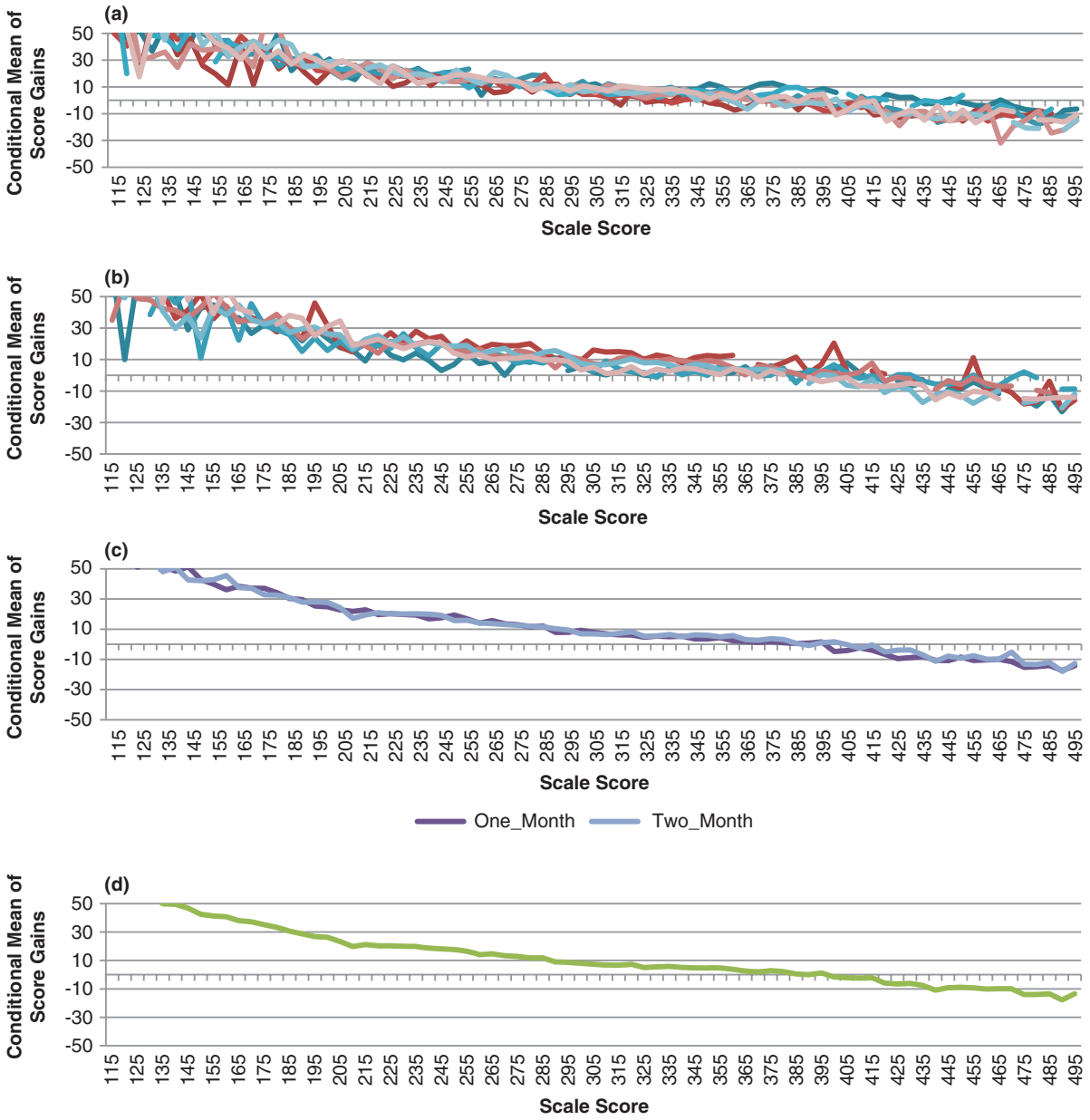

Figure 1 Conditional means of score gains for Listening: (a) 1-month gap, (b) 2-month gap, (c) 1- and 2-month gaps, and (d) 1- or 2-month gap.

over a 2-month interval between adjacent administrations. Therefore, we decided to use the conditional distribution based on combined and accumulated data for SGGM scaling for Listening.

Reading

Table 7 shows the summary statistics of repeaters' Reading scale score gains. As the table shows, for some repeater groups, the average score gains were negative. For the eight pairs of administrations with a 1-month gap, the average score gains ranged from −2.74 to 11.58, with standard deviations from 42.46 to 45.16. Combining the eight pairs of adjacent administrations, the average score gain was 6.26, with a standard deviation of 43.73. For the six pairs of administrations with a 2-month gap, the average score gains ranged from −0.03 to 17.47, with standard deviations from 41.35 to 45.80. Combining the six pairs of adjacent administrations, the average score gain was 6.96, with a

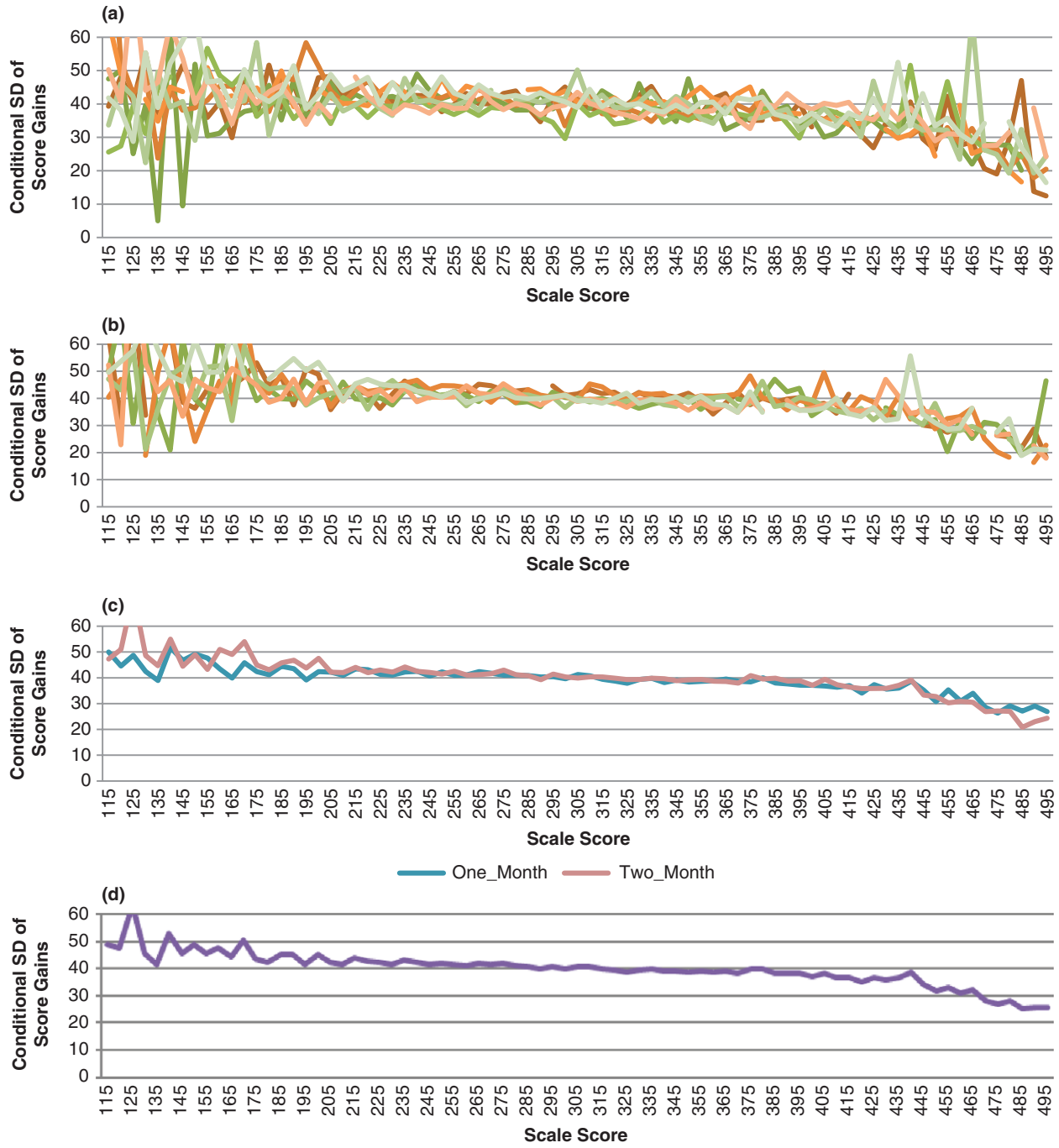

Figure 2 Conditional standard deviations of score gains for Listening: (a) 1-month gap, (b) 2-month gap, (c) 1- and 2-month gaps, and (d) 1- or 2-month gap.

standard deviation of 44.39. Among the 14 pairs of administrations, the standardized mean score changes varied from −0.03 to 0.20, with an average of 0.08 based on accumulated data. Therefore, there was variability in the average score gains across different pairs of adjacent administrations. As we found in the listening section, there was considerable variability in examinees' score gains, but the dispersion of score gains was very consistent across different pairs of adjacent administrations. Both the average and the variability of score gains of the accumulated 1-month repeaters and 2-month repeaters were very similar, and their combined average score gain was 6.61, with a standard deviation of 44.07.

Table 7 Summary Statistics of Repeaters' Reading Scale Score Gains

			Earlier admin.		Later admin.		Score gain		
Time gap	Adjacent administration	Sample size	*M*	*SD*	*M*	*SD*	*M*	*SD*	Standardized difference
1 month	May 10 – June 10	9,188	270.49	92.20	267.74	90.15	−2.74	45.16	−0.03
	June 10 – July 10	6,251	269.40	91.08	277.00	86.34	7.59	44.16	0.09
	September 10 – October 10	12,666	274.97	86.55	280.30	87.00	5.33	43.53	0.06
	October 10 – November 10	14,598	272.12	84.32	280.70	89.76	8.58	42.84	0.10
	May 11 – June 11	11,972	270.22	83.41	278.13	87.80	7.92	43.92	0.09
	June 11 – July 11	7,935	278.45	90.20	276.65	85.96	−1.80	43.31	−0.02
	September 11 – October 11	15,334	276.58	85.30	288.16	86.30	11.58	43.68	0.13
	October 11 – November 11	15,920	279.73	84.68	286.89	87.83	7.16	42.46	0.08
	Accumulative	93,864	274.47	86.57	280.73	87.94	6.26	43.73	0.07
2 months	July 10 – September 10	9,979	272.60	84.98	276.90	89.01	4.30	44.39	0.05
	November 10 – January 11	14,405	270.08	89.60	281.97	86.63	11.89	42.94	0.13
	July 11 – September 11	13,033	273.27	85.74	276.83	88.75	3.56	44.41	0.04
	November 11 – January 12	16,971	277.24	87.58	277.21	88.33	−0.03	41.35	0.00
	January 12 – March 12	19,735	268.67	87.99	286.14	89.16	17.47	44.06	0.20
	March 12 – May 12	22,491	275.86	87.36	278.88	86.77	3.02	45.80	0.03
	Accumulative	96,614	273.09	87.46	280.05	88.08	6.96	44.39	0.08
1 or 2 months	Accumulative	190,478	273.77	87.02	280.38	88.01	6.61	44.07	0.08

We also examined the conditional means and standard deviations of scale score gains at each scale score level for the reading section. Figure 3 shows the conditional means of score gains, and Figure 4 shows the conditional standard deviations of score gains based on different data sets. Again, the conditional statistics for scale scores below 90, which approximately corresponds to the chance raw score level of 25, were not presented in the figures because very few examinees had scores in that range.

As found in the listening section, the score gains for the reading section were also different across different scale score points. For examinees with higher earlier scale scores, scale scores tended to drop. For examinees with lower earlier scores, scores tended to increase. As seen in Figures 4c and d, the variability of score gains was also related to repeaters' earlier score levels: The dispersion of individual score gains tended to be relatively smaller at the higher score levels. However, Figures 4a and b show relatively bigger variability across pairs of administrations. This larger across-pair variability is likely due to the comparatively smaller number of examinees with higher scores in the earlier administrations. At each scale score point, there was some variability in both conditional means and standard deviations of score gains across the different pairs of adjacent administrations with either 1- or 2-month gaps. Compared with Listening, except at lower scale score levels, the conditional means of score gains were more inconsistent across different pairs of administrations. The conditional standard deviations of score gains were larger at the top level but smaller at the lower level. The larger conditional standard deviations at the higher scale scores may be related to the relatively smaller frequencies at the top score levels for Reading. However, the conditional statistics of score gains were relatively consistent between accumulated 1-month-gap and 2-month-gap administrations, except for the conditional standard deviations at the top score levels. Therefore, we decided to use the conditional distribution based on combined and accumulated data for SGGM scaling for Reading.

Conversion Differences

We used raw-to-scale conversion difference plots (i.e., SGGM scaling conversion minus NEAT equating conversion) to illustrate the conversion differences between the two linking methods (see Figure 5 for Listening and Figure 6 for Reading). The plot above 0 indicates that the SGGM conversion was higher than the NEAT conversion, and the plot below 0 suggests that the SGGM conversion was lower than the NEAT conversion. In the graphs, the conversions were truncated at chance level, which is the raw score 28 for Listening and the raw score 25 for Reading, and very few examinees scored at or lower than those levels. At the top of the scale, all scale scores at and higher than 495 were truncated to 495, which is the maximum of reported scale scores.

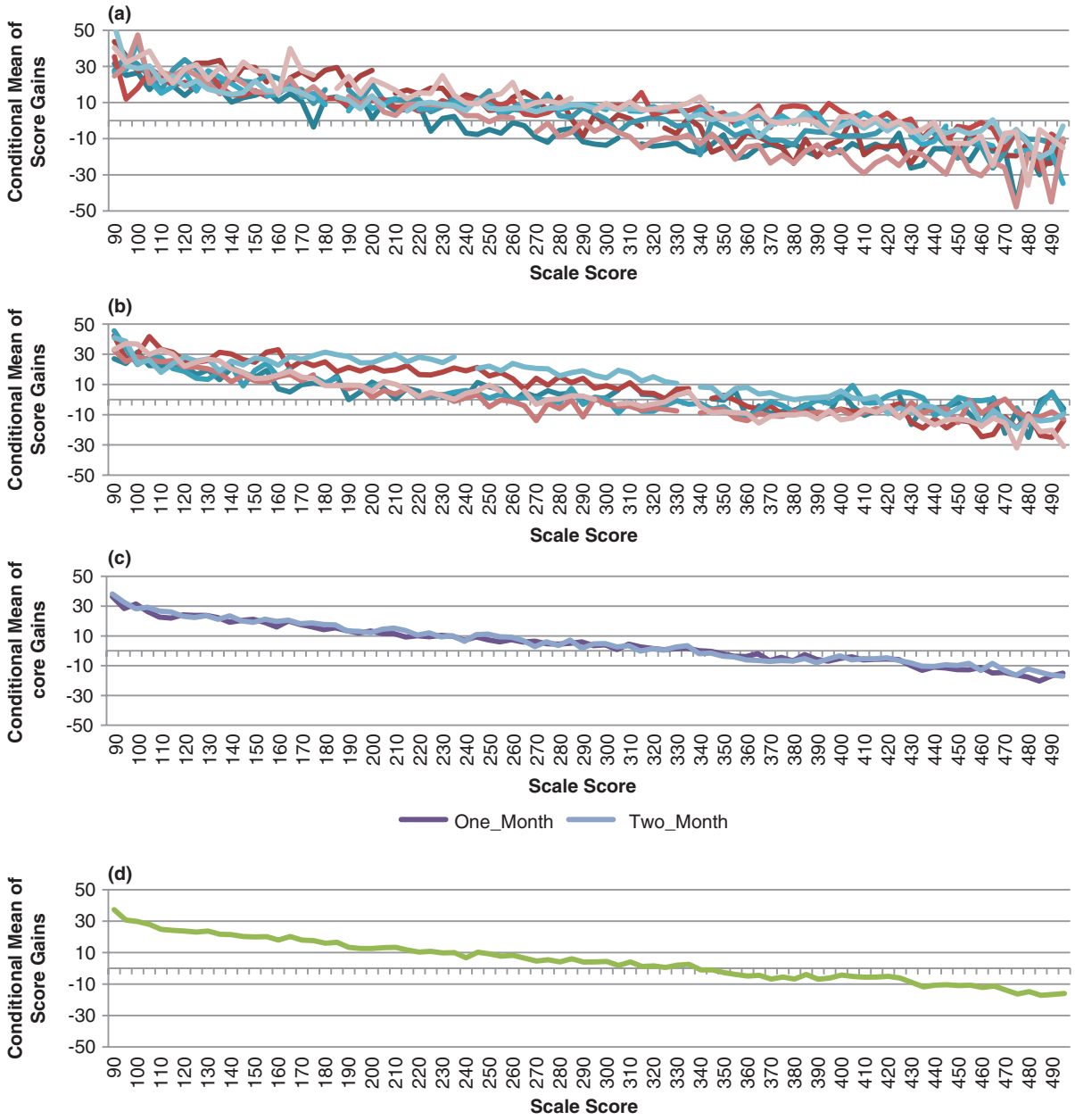

Figure 3 Conditional means of score gains for Reading: (a) 1-month gap, (b) 2-month gap, (c) 1- and 2-month gaps, and (d) 1- or 2-month gap.

Listening

Figure 5 shows the conversion difference plots for the listening section for the 10 forms. For Test Forms A, E, and F, the SGGM scaling tended to produce higher conversions than the NEAT equating, except in some small score ranges. For Test Forms C, D, and H, the SGGM scaling tended to produce lower scale scores in all or most score ranges. For the remaining four forms, the SGGM scaling typically produced lower conversions in the lower score range (i.e., lower than the raw score of 50), higher conversions in the middle score range (i.e., the raw scores of 55–85), and then lower conversions in the top score range (i.e., the truncated raw score points). Strictly speaking, for all of the 10 test forms, no SGGM scaling function was the same as the NEAT equating function, and the conversion differences could reach to as low as −20 scale score points or as high as 10 scale score points, which are approximately 1/4 and 1/8 of a standard deviation, respectively.

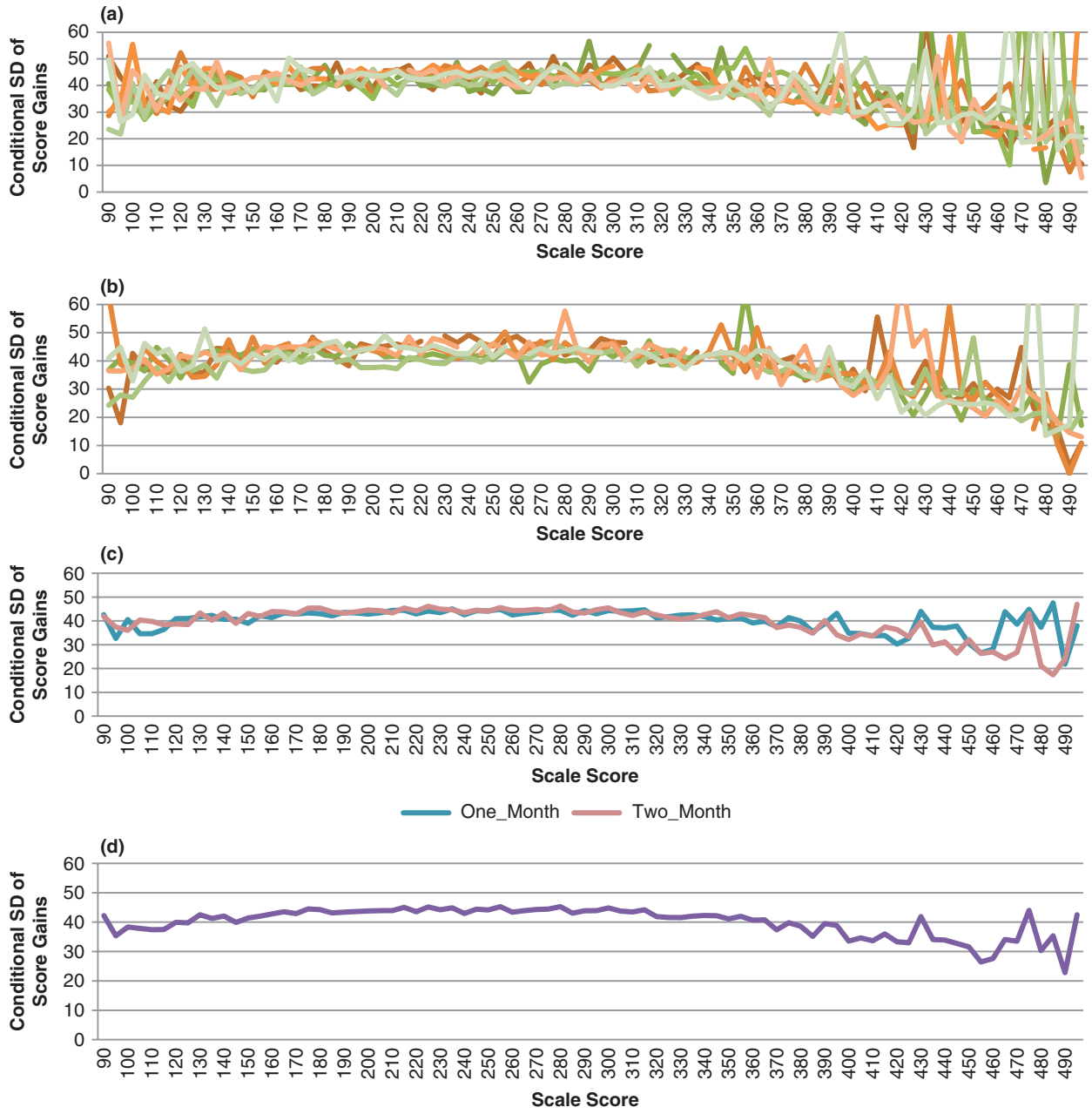

Figure 4 Conditional standard deviations of score gains for Reading: (a) 1-month gap, (b) 2-month gap, (c) 1- and 2-month gaps, and (d) 1- or 2-month gap.

Reading

Figure 6 shows the conversion difference plots for the reading section for the 10 forms. Based on the plots, for Test Forms A, C, G, and I, the SGGM scaling tended to produce higher conversions than the NEAT equating, except in the very low score range. For Test Form D, the SGGM scaling produced lower scale scores in all score ranges. For the remaining five forms, the SGGM scaling produced lower conversions in some score ranges but higher conversions in the other score ranges, depending on specific test forms. As we found in the listening section, for all of the 10 test forms, no SGGM scaling function was the same as the NEAT equating function, and the conversion differences could reach to as low as −20 scale score points or as high as 17 scale score points, which are approximately 1/4 and 1/5 of a standard deviation, respectively.

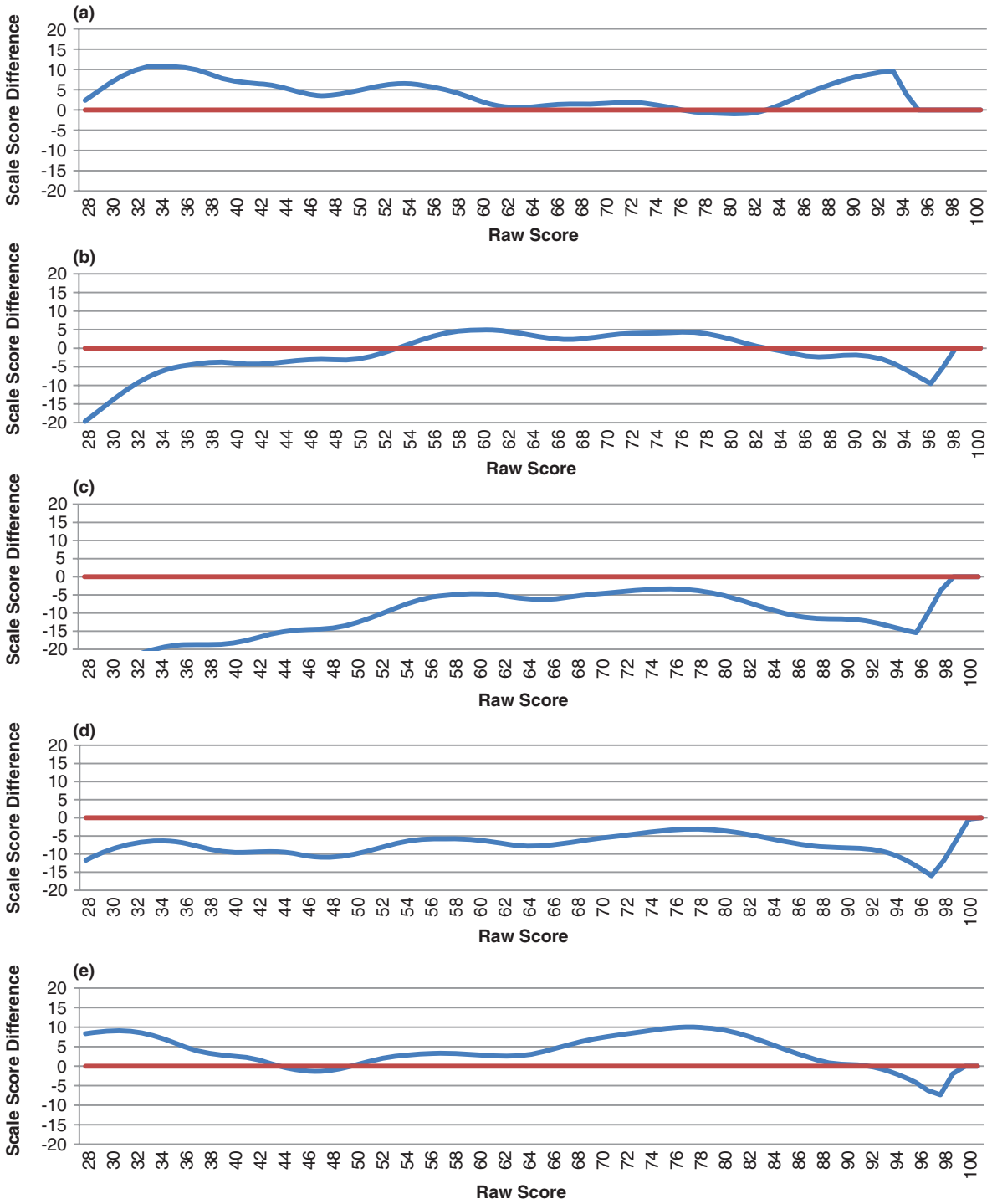

Figure 5 Conversion differences (SGGM-NEAT) for Listening: (a) Test Form A, (b) Test Form B, (c) Test Form C, (d) Test Form D, (e) Test Form E, (f) Test Form F, (g) Test Form G, (h) Test Form H, (i) Test Form I, and (j) Test Form J.

Figures 7 and 8 show the RMSD for the listening and reading sections, respectively. Based on the two plots, the RMSD tended to be greater than 3 but less than 12 for both Listening and Reading across the score ranges, with relatively larger values at the top score ranges and smaller values in the middle score ranges.

Figure 9 shows the WRMSD for both the listening and reading sections for the 10 test forms. Based on the figure, the WRMSD varied from 3 to 9 scale score points for Listening and from 3 to 6 scale score points for Reading.

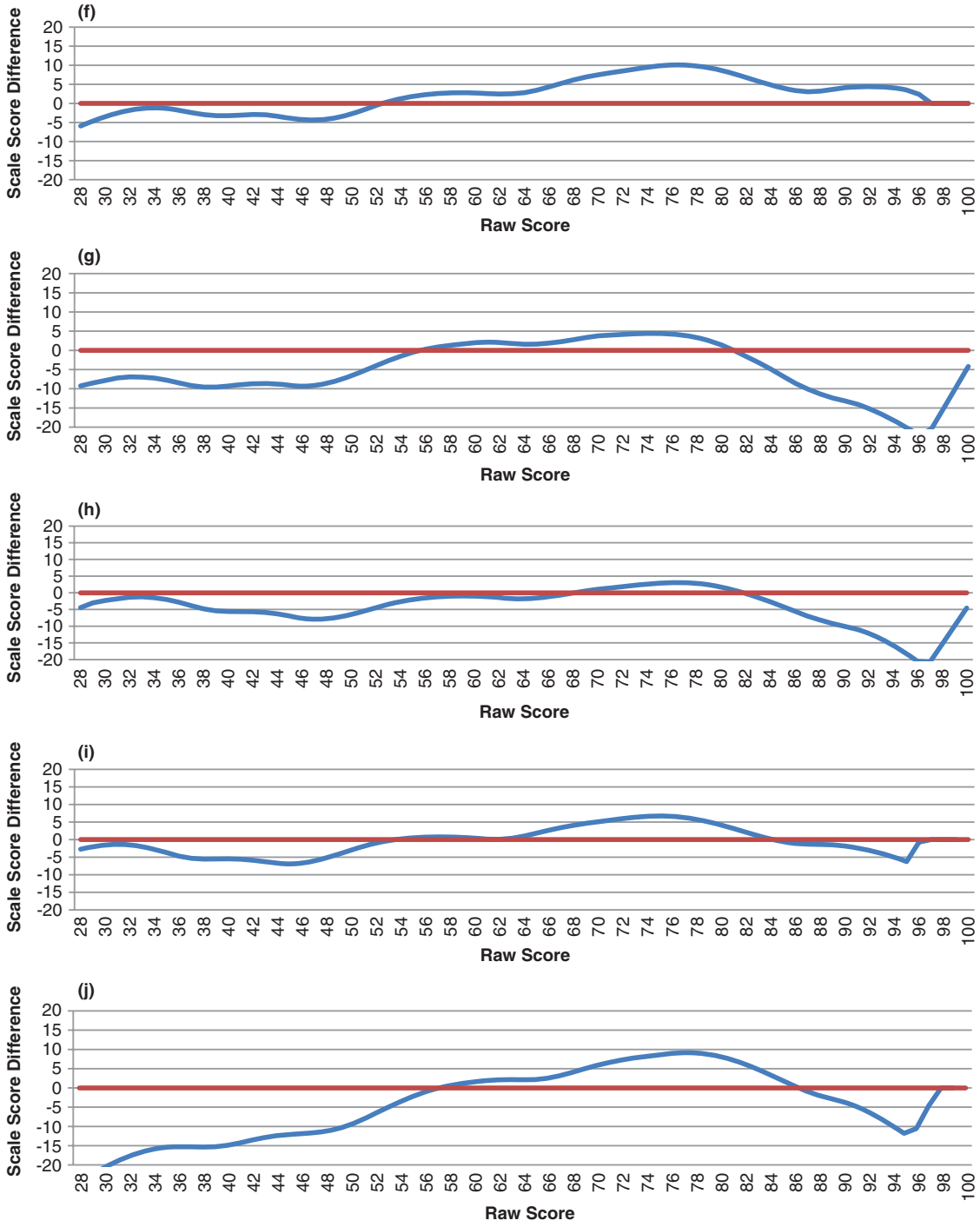

Figure 5 Continued.

Summary Statistics

After applying the raw-to-scale score conversions from the SGGM scaling method to all examinees in each administration, we obtained their scale scores and their summary statistics (i.e., means and standard deviations). Table 8 displays the summary statistics that resulted from the SGGM scaling conversions and NEAT equating conversions for Listening and Reading for each of the 10 forms. The table also shows the differences in the summary statistics between the two methods. Consistent with what we found in the conversion difference plots, the scale score means based on the SGGM scaling

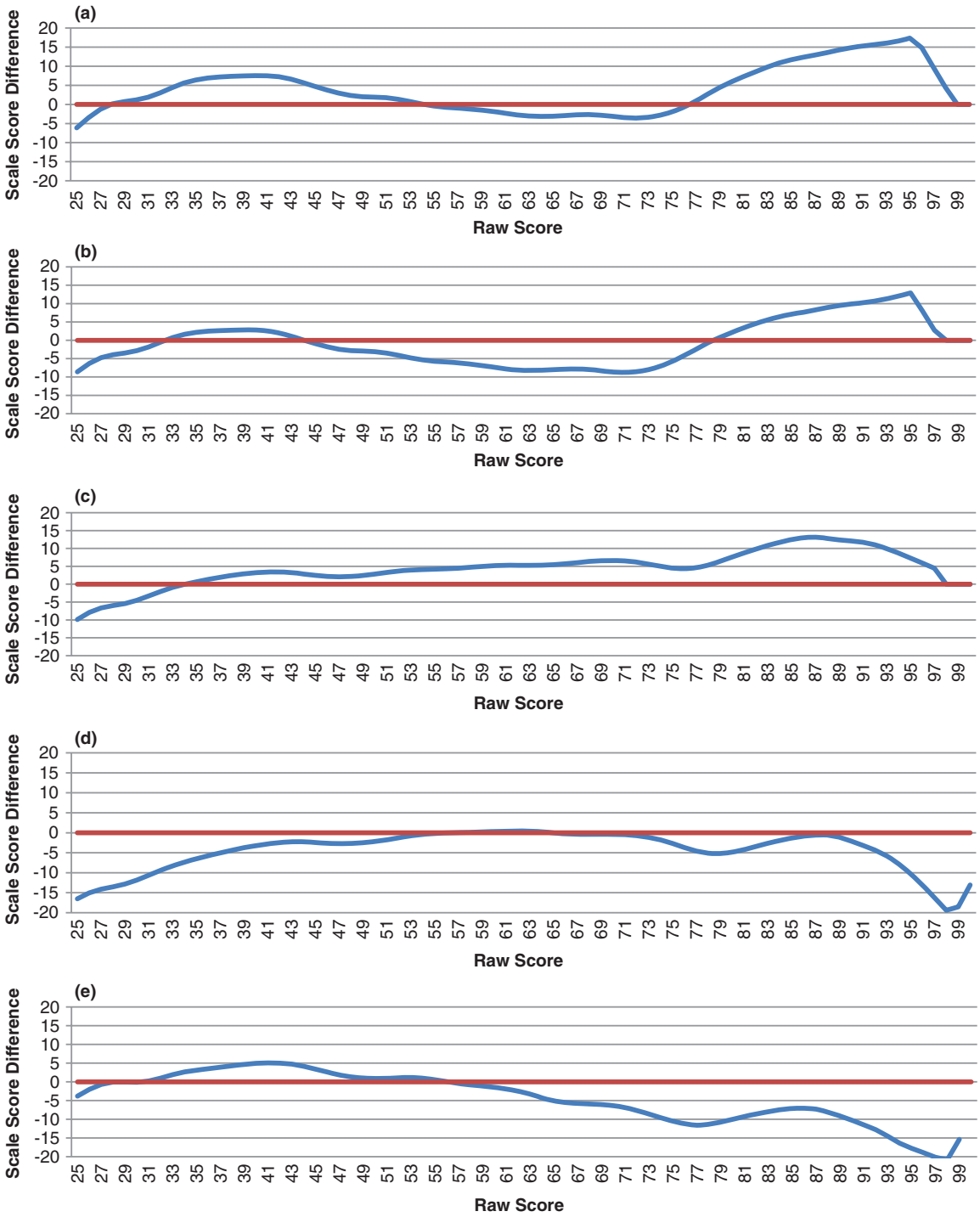

Figure 6 Conversion differences (SGGM-NEAT) for Reading: (a) Test Form A, (b) Test Form B, (c) Test Form C, (d) Test Form D, (e) Test Form E, (f) Test Form F, (g) Test Form G, (h) Test Form H, (i) Test Form I, and (j) Test Form J.

method were sometimes higher and sometimes lower than those based on the NEAT equating method, depending on specific sections and test forms. Specifically, compared with NEAT equating results, the SGGM scaling produced from 7.72 lower average scale score points to 3.64 higher average scale score points for Listening and from 3.19 lower average scale score points to 4.81 higher average scale score points for Reading. However, the average signed difference between the means was −0.66 for Listening and 0.48 for Reading. The average signed difference for the standard deviation was 0.72

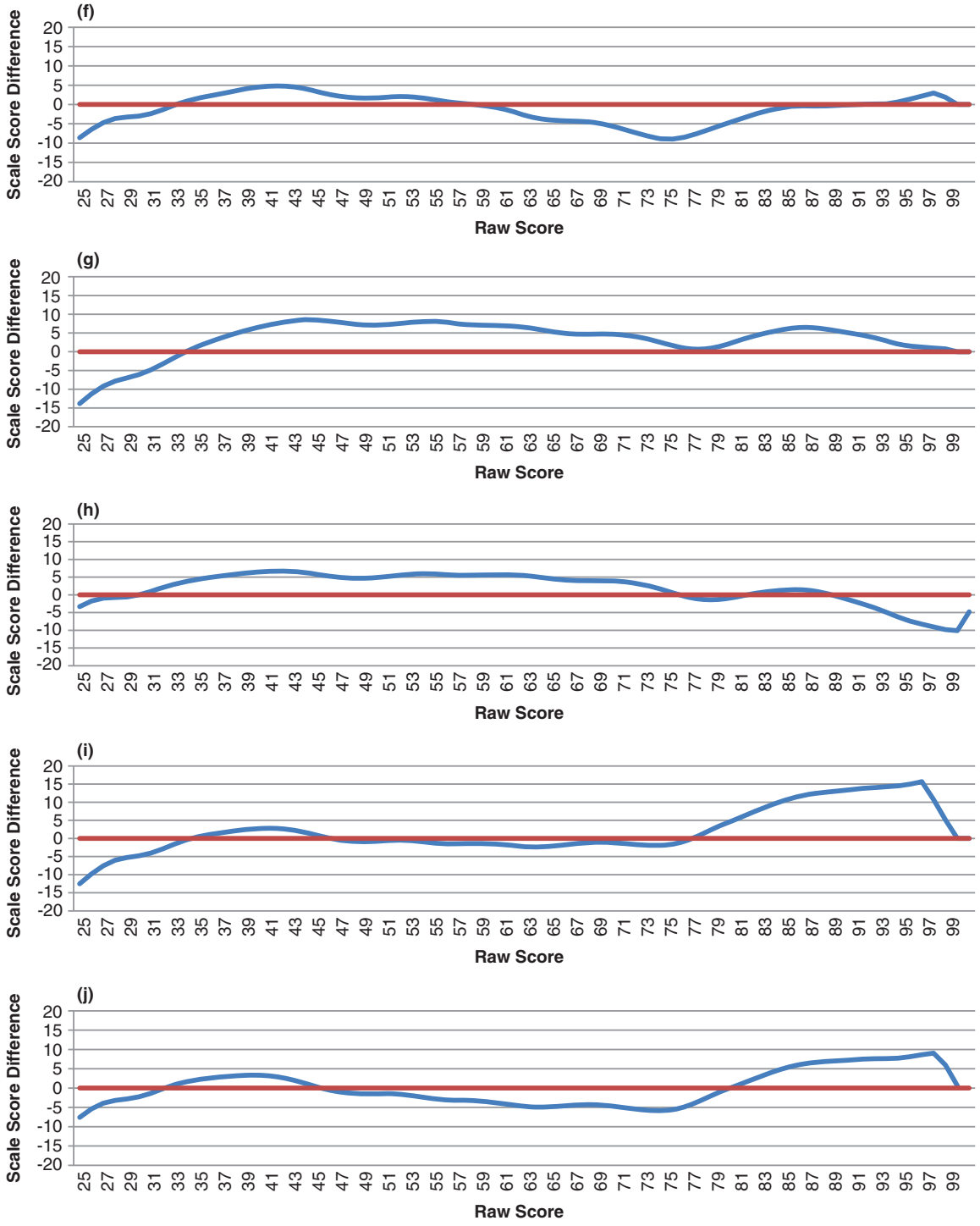

Figure 6 Continued.

for Listening and −0.17 for Reading. The average of the absolute differences in the scale score mean was 3.20 for Listening and 2.51 for Reading. The average of the absolute differences in the standard deviation of the scale scores was 1.57 for Listening and 1.55 for Reading. Therefore, at the level of average score, SGGM scaling did not show a systematic bias of higher or lower average scores. Neither were the distributions of scale scores systematically more or less variable than the NEAT equating results.

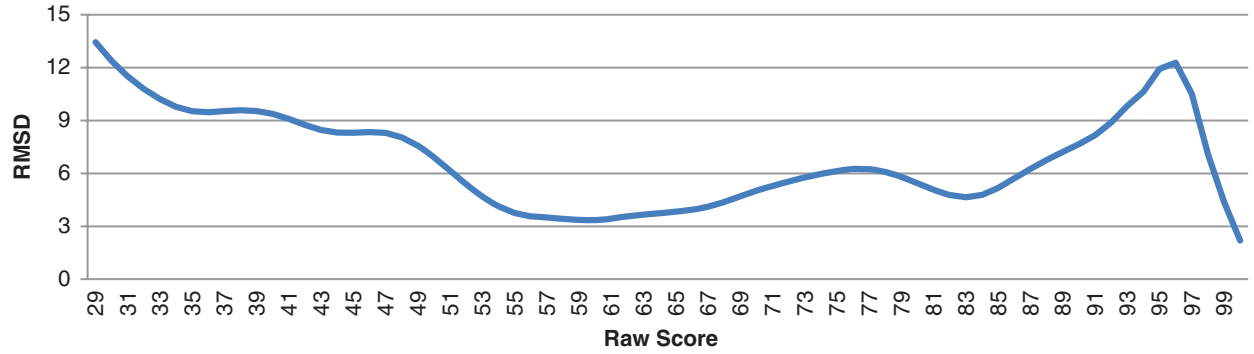

Figure 7 Root-mean-square of differences (RMSD) between single-group growth model (SGGM) and nonequivalent groups with anchor test (NEAT) conversions for Listening.

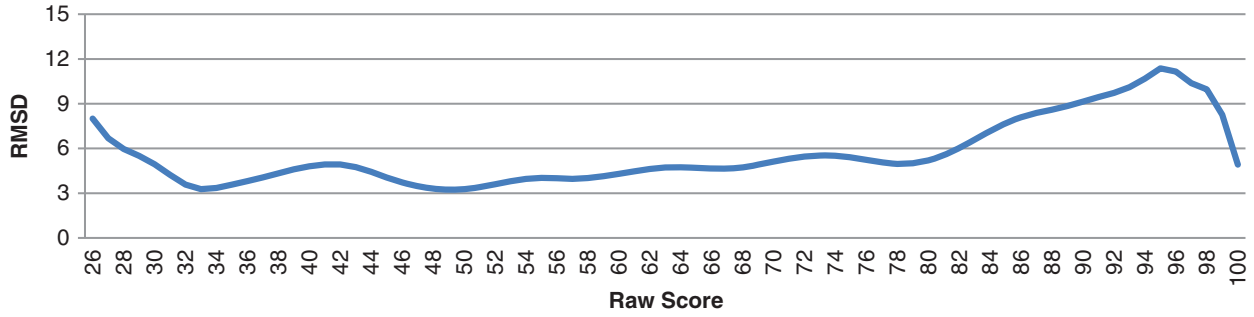

Figure 8 Root-mean-square of differences (RMSD) between single-group growth model (SGGM) and nonequivalent groups with anchor test (NEAT) conversions for Reading.

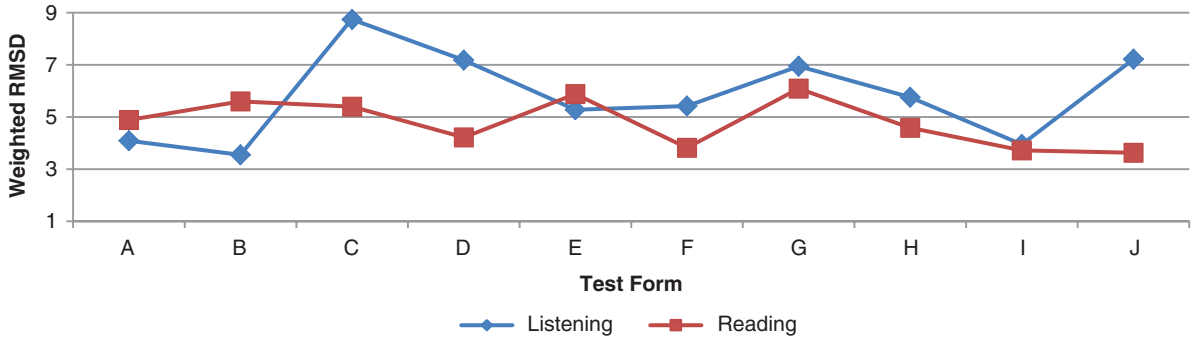

Figure 9 Weighted root-mean-square of differences (WRMSD) between single-group growth model (SGGM) and nonequivalent groups with anchor test (NEAT) conversions for Listening and Reading.

Discussion

All linking methods have their own assumptions, and SGGM scaling is no exception. SGGM scaling assumes that the adjacent-administration repeaters' conditional distribution of later scale scores, given earlier scale scores, is consistent across different pairs of adjacent administrations. In other words, it assumes that the conditional distributions of repeaters' scale score gains are consistent across administrations. Based on the eight pairs of adjacent administrations with a 1-month gap and six pairs with a 2-month gap, the means of repeaters' score gains were small, with the average score gain of 7 (i.e., 1/12 of a standard deviation) for both Listening and Reading, but they differed somewhat across different pairs of adjacent administrations. On average, the standard deviations of score gains were 41 for Listening and 44 for Reading, and they were very consistent across different pairs of administrations. This finding of smaller average score gains but large variability of score gains was consistent with the results from a repeater analysis study using graduate admissions

Table 8 Summary Statistics Comparison Between Nonequivalent Groups With Anchor Test (NEAT) Equating and Single-Group Growth Model Scaling (SGGM)

		NEAT		SGGM		Difference	
Form	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Listening							
A	36,126	315.26	84.43	318.49	82.86	3.23	−1.57
B	33,126	314.41	83.37	315.53	83.72	1.12	0.35
C	42,721	321.24	85.98	313.52	87.32	−7.72	1.34
D	40,233	320.74	86.53	314.05	86.95	−6.70	0.42
E	40,732	314.00	86.66	317.61	87.04	3.60	0.38
F	41,197	312.78	84.75	316.43	87.56	3.64	2.81
G	32,140	322.38	85.63	320.12	84.18	−2.26	−1.45
H	32,283	325.03	84.53	322.47	83.27	−2.56	−1.26
I	37,574	315.77	85.74	316.88	87.52	1.11	1.78
J	59,958	313.85	85.60	313.76	89.97	−0.09	4.37
Reading							
A	36,126	262.28	92.19	264.11	91.98	1.84	−0.21
B	33,126	264.03	92.78	260.84	92.07	−3.19	−0.71
C	42,721	259.17	91.35	263.49	94.69	4.32	3.34
D	40,233	266.65	94.36	264.27	95.65	−2.38	1.28
E	40,732	261.28	95.08	258.66	90.81	−2.62	−4.27
F	41,197	255.56	92.69	255.00	91.14	−0.56	−1.55
G	32,140	263.48	92.36	268.29	92.89	4.81	0.53
H	32,283	267.36	94.43	271.20	92.92	3.85	−1.50
I	37,574	259.87	91.75	259.99	93.46	0.12	1.72
J	59,958	258.62	94.31	257.18	93.97	−1.44	−0.34

exam data (Yang, Bontya, & Moses, 2011). The 1-month repeaters' standardized average score gains of 0.08 for Listening and 0.07 for Reading were relatively smaller than the results from 1-month repeater analysis based on data from the *TOEFL*® test (i.e., 0.12 for Listening and 0.17 for Reading; see Zhang, 2008). The score gains across the scale score levels show that both the conditional means and standard deviations of score gains were not consistent across different pairs of adjacent administrations, especially at lower and/or higher ends of the scale. Based on the accumulated data across administrations, the score gains for 1-month repeaters were similar to those for 2-month repeaters. It seems reasonable to use the conditional distributions based on combined 1-month and 2-month data for scaling, although the conditional distributions across individual pairs of adjacent administrations were not very consistent.

Not surprisingly, there was a regression effect for repeaters' score changes, which is due to less than perfect test score reliability (i.e., 0.92 for Listening and 0.93 for Reading). Specifically, for the examinees with higher scale scores in earlier administrations, scale scores in later administrations tended to be lower; for examinees with lower scores in earlier administrations, scores in later administrations tended to be higher. Yang et al. (2011) also found this pattern of score changes on a graduate admissions exam. It was also not surprising that the 1-month repeaters had slightly less score gain than the 2-month repeaters (i.e., 0.08 and 0.07 vs. 0.10 and 0.08 for Listening and Reading, respectively), although some repeater groups' reading score means dropped slightly within 1 month.

The larger volume based on combined 1-month and 2-month repeaters served to make the conditional distribution more stable. However, the assumed invariance in conditional distributions (i.e., score gains) across different administrations would certainly affect the accuracy of SGGM scaling if score gain differs across administrations. Differences were found between NEAT conversions and SGGM conversions. However, the bias in the conversion differences and summary statistics was not consistent. Compared with the NEAT conversions, the SGGM conversions were sometimes higher and sometimes lower, depending on test sections and forms. Similarly, examinees' scale score means and standard deviations based on SGGM scaling were sometimes higher and sometimes lower than with the NEAT equating conversions.

No equating method is perfect, and all equating methods may have their own problems in the real world of operational work. Therefore, any criterion used for evaluation is not the absolute truth. The NEAT design equating used for the

current study typically worked well. There were small ability differences between the new and reference groups, along with consistent anchor item performance in the new and reference test forms. A scale stability study (Lu, Haberman, & Liu, 2013) examined the variations in the means and standard deviations of scale scores and the variations in the raw-to-scale conversions across 61 forms in the testing program and found that the scale of the test scores was fairly stable across administrations. Based on the findings of the current study, the average signed mean differences of −0.66 for Listening and 0.48 for Reading and the average signed standard deviation differences of 0.72 for Listening and −0.17 for Reading between the two linking methods were very small. Therefore, the SGGM scaling method did not show a systematic bias of higher or lower average scores. Neither were the distributions of scale scores systematically more or less variable than the NEAT equating results. This conclusion is consistent with the findings from a prior study (Wei, 2013) that used single-group equating as the criterion to evaluate the SGGM scaling function.

A close look at the conversion difference plots reveals that the SGGM conversions are more different from the NEAT conversions at the top and bottom score ranges than at the middle score range. This pattern is consistent with the finding that both the conditional means and the standard deviations of score gains vary the most across different administrations at lower and/or higher ends of the scale than at the middle of the scale. The bigger differences at the bottom and top score ranges might come from both the conditional distributions in SGGM scaling and from random error in the NEAT equating because of smaller sample sizes in those ranges.

The conversion difference plots in Figures 5 and 6 also show that there are abrupt jumps or drops in the top score ranges for both the listening and reading sections in most of the test forms. This may be related to the different top score ranges used in the two linking methods. For SGGM linking, we used the truncated scale scores from historical data to compute the conditional distribution. Then we used the new-form repeaters' earlier truncated scale scores to estimate their scale scores in the new administration. We then conducted equipercentile scaling between new repeaters' observed raw scores and their estimated and truncated scale scores to create the conversions for the new test form. Therefore, all the steps of the SGGM scaling used truncated scale scores. For NEAT equating, we first equated the new form's raw scores to the old form's raw scores using common items then put them on the reference scale through the old form's raw-to-scale conversion. We did not truncate scale scores in NEAT equating until the last step for score reporting. This difference between the two linking methods might have a consistent impact on their conversion differences on the top score range.

Some other factors might also have an impact on the observed differences between SGGM scaling results and NEAT equating results, and more effort needs to be made to improve the SGGM scaling function. For example, the conditional distribution used for scaling in this study did not take into consideration the number of prior administrations the repeaters had taken before the later administration. Based on operational data, the first-time repeaters showed somewhat more improvement than those who had taken the test more frequently. This finding suggests that different repeaters may have different score gain patterns. The SGGM scaling function would improve if the conditional distribution were to include such information. Cleaning repeaters' data by removing outliers or unusual scores would also improve SGGM scaling. For example, we found in the operational data that some examinees had score gains or drops of more than 400, which is almost impossible, unless either their earlier scores or later scores did not really reflect their ability levels. We should remove these repeaters from the data before we estimate the conditional distribution in the future.

We should also take into account the characteristics of the criterion when we evaluate the SGGM scaling function. No equating method is perfect, and all equating methods may have their own problems in the real world. Therefore, any criterion used for evaluation is not the absolute truth. The NEAT equating used in the current study typically worked well based on test score equating guidelines and suggestions (Kolen & Brennan, 2004; Livingston, 2014). For example, the new and reference groups came from the same population, and their ability differences were typically small, with the standardized group differences being less than 0.1 and the ratio of variances being close to 1.0. The total test and the common-item sets used the same content and statistical specifications. The security of common-item sets was closely monitored by comparing the average p values of the common-item sets and other items in the test. Difficulty estimate (i.e., percentage correct or p value) plots of the common items based on new and reference groups suggest that the common items performed consistently in the new and reference groups, with correlations being always larger than 0.97. Therefore, the authors believe that the NEAT equating procedure is satisfactory and the equating results can be used to evaluate the reasonableness of other methods (e.g., Haberman, 2015). However, some inevitable equating errors likely resulted from either

sampling or violation of equating assumptions. We should consider the range of errors while evaluating the SGGM scaling results.

The SGGM scaling method does not require common items between test forms, and we can examine its assumptions using historical data. However, the SGGM method obtains advantages at the cost of limitations. For example, there should be large volumes of identified repeaters between adjacent administrations, and their prior reported scale scores should be available. In this study, two types of historical data were required to implement the SGGM method. First, before SGGM scaling, the repeaters' historical data from May 2010 to May 2012 had to be used to set up and examine the conditional distribution of repeaters' later scale scores, given their earlier scale scores. Second, in the SGGM scaling for the 10 new forms administered from June to November 2012, the repeaters' scale scores in the earlier administrations, which were used to estimate repeaters' scale scores in later administrations, were already available based on NEAT design equating. As a result, the accuracy of the SGGM scaling for the new forms depends in part on the quality of NEAT design equating in previous administrations. One follow-up question should address what would happen if we used the repeaters' scale scores in the earlier administration from SGGM scaling to estimate their scale scores in the later administration and conducted the SGGM scaling sequentially. To evaluate the feasibility of totally replacing NEAT equating by SGGM scaling for all the new forms, except for the two starting forms, SGGM scaling results would be used to estimate repeaters' scale scores in later administrations. In this way, we could examine the long-term impact or chained effect of the SGGM method. Therefore, future studies should only use scale scores from SGGM scaling to evaluate the long-term performance of the method for the testing program.

Conclusion

We evaluated the assumption and scaling results of the SGGM method in this study using the results from NEAT equating as the criterion. Based on the repeat examinees' data from 14 pairs of adjacent administrations, the conditional distributions of repeaters' later scale scores on their earlier scale scores were not consistent, but the 1-month conditional distributions were very similar to the 2-month conditional distributions. Based on the scaling results of 10 test forms, the SGGM scaling conversions were different from the NEAT equating conversions and had an impact on examinees' scoring results. However, compared with examinees' summary statistics based on NEAT equating, the SGGM scaling did not show a systematic bias in either the average or the variability of scale scores. More studies need to be conducted to examine the long-term impact of replacing NEAT equating by SGGM scaling in a testing program.

Acknowledgments

The authors thank Samuel Livingston, Gautam Puhan, and Hongwen Guo for their comments and suggestions on earlier versions of this report. The authors also thank Matthew Shotts, Albert Low, Hexin Chang, Jeffrey Kaufman, and Keith Segreto for their contributions to the data management and analyses of this study.

References

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40,* 254 – 273.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133 – 183.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practice* (2nd ed.). New York, NY: Springer.

Liao, C. W., & Livingston, S. A. (2012, April). *A search for alternatives to common-item equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.

Livingston, S. A. (2005). *Using repeater data to equate scores.* Unpublished manuscript.

Livingston, S. A. (2014). *Equating test scores (without IRT)* (2nd ed.). Princeton, NJ: Educational Testing Service.

Lu, R., Haberman, S., & Liu, J. (2013). *Scale stability check of a large scale language test.* Unpublished manuscript.

Morgan, R. (2011). *Equating through common examinees: The repeater model solution.* Unpublished manuscript.

Wei, Y. (2013). *Evaluation of common-item cross-population linking and single group growth model scaling for a language test.* Unpublished manuscript.

Wei, Y., Liu, J., & Dorans, N. (2013). *Evaluation and exploration of linking design for a language test*. Unpublished manuscript.

Yang, W. L., Bontya, A. M., & Moses, T. P. (2011). *Repeater effects on score equating for a Graduate Admission Exam* (Research Report No. RR-11-17). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02253.x

Zhang, Y. (2008). *Repeater analysis for TOEFL iBT* (Research Memorandum No. RM-08-05). Princeton, NJ: Educational Testing Service.

Suggested citation:

Wei, Y., & Morgan, R. (2016). *An evaluation of the single-group growth model (SGGM) as an alternative to common-item equating* (ETS Research Report No. RR-16-01). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12087

Action Editor: Guatam Puhan

Reviewers: Samuel Livingston and Hongwen Guo